Adversarial Training of Variational Auto-encoders for High Fidelity Image Generation

Salman H. Khan†, Munawar Hayat ‡, Nick Barnes †
†Data61 - CSIRO and ANU, Australia, ‡University of Canberra, Australia,
{salman.khan,nick.barnes}@data61.csiro.au, munawar.hayat@canberra.edu.au

Abstract

Variational auto-encoders (VAEs) provide an attractive solution to image generation problem. However, they tend to produce blurred and over-smoothed images due to their dependence on pixel-wise reconstruction loss. This paper introduces a new approach to alleviate this problem in the VAE based generative models. Our model simultaneously learns to match the data, reconstruction loss and the latent distributions of real and fake images to improve the quality of generated samples. To compute the loss distributions, we introduce an auto-encoder based discriminator model which allows an adversarial learning procedure. The discriminator in our model also provides perceptual guidance to the VAE by matching the learned similarity metric of the real and fake samples in the latent space. To stabilize the overall training process, our model uses an error feedback approach to maintain the equilibrium between competing networks in the model. Our experiments show that the generated samples from our proposed model exhibit a diverse set of attributes and facial expressions and scale up to high-resolution images very well.

1. Introduction

Recent advances in deep learning have seen significant success in discriminative modeling for a wide range of classification tasks [17, 16, 10, 9, 15]. Generative models, however, still face many challenges in modeling complex data in the form of images and videos. Despite being fairly challenging, generative modeling of images is desirable in many applications. These include unsupervised and semi-supervised feature learning from large-scale visual data [25], understanding the representations learned by the discriminative models [32], image completion [31], denoising, super-resolution [20] and prediction of future frames in a video [30]. Auto-encoder based models have traditionally been used for the generative modeling task. Variational Auto-Encoders (VAEs) are their improved variants which restrict the learned latent space representation to a prior probability distribution [18]. The VAE based models approximate the data likelihood very well, however, their generated images are of low quality, do not retain fine details and have a limited diversity.

Generative adversarial networks (GANs) provide a viable solution to the low quality output from VAEs [8]. In theory, GANs can more accurately estimate the data distribution given an infinite amount of data and generate more realistic images. However in practice, GANs are difficult to optimize due to the lack of a closed-form loss function and can generate visually absurd outputs [1]. Wasserstein metric [2] and energy based [33] adversarial variants have been proposed to reduce the instability of GANs. A common limitation of all these approaches is the lack of control over the latent representations and therefore making it difficult to generate data with the desired attributes. To resolve this issue, [19] introduced an adversarial loss function to train VAEs. Though this helps stabilize the training and gives control over the latent distribution, the generated images are not sharp and crisp enough compared to their counterpart GAN based approaches.

In this paper, we propose to match the data as well as the reconstruction loss and latent distributions for real and fake samples in the VAE during the training process. This allows us to recover high quality image samples while having full control over the latent representations. Leveraging on the learned latent space for both real and fake images, we introduce priors to ensure the generation of perceptually plausible outputs. The complete model is trained using an adversarial loss function which allows us to learn a rich similarity measure for images. This leads to a highly flexible and robust VAE architecture, that combines the benefits of both variational and adversarial generative models. Furthermore, drawing insight from the recent boundary equilibrium GAN [4], we design a controller to balance the game between the generator and the discriminator. This results in a smooth optimization and allows us to avoid the commonly employed heuristics for stable training.

In summary, this paper makes the following contribu-

tions: 1) We propose a new VAE+GAN model, capable of generating high fidelity images. The proposed model encodes images in a compact and meaningful latent space where different arithmetic operations can be performed and reflected back in the image domain. 2) Our proposed model incorporates a learned similarity metric to avoid unrealistic outputs and generates globally coherent images which are perceptually more appealing. 3) To stabilize the model training, we propose to consider the past and future trends of the error signal obtained by comparing the discriminator and generator losses.

Next, we outline the related work followed by our proposed model.

2. Related Work

Generative image modeling using dictionary learning techniques have been extensively studied in the literature with applications to texture synthesis [6], in-painting [11] and image super-resolution [7]. In comparison, generating natural images did not see much success until recently with advances in deep learning. Restricted Boltzmann machines [13] and deep auto-encoder models [12] are amongst the earliest neural networks based methods for unsupervised feature learning and image generation. These models first encode images into a latent space and then decode them back into image space by minimizing pixel-wise reconstruction errors. Kingma and Welling [18] proposed a variational inference based encoding-decoding approach which enforces a prior on the learnt latent space. Their proposed method achieved promising results but the generated images were often blurry.

Auto-encoder based models are trained to minimize pixel-level reconstruction errors. These models do not consider holistic contents in an image as interpreted by human visual perception. For example, a small scale rotation would yield large pixel-wise reconstruction error, but would be barely noticeable to human visual perception. Generative Adversarial Networks [8] can learn a better similarity metric for images, and have received significant research attention since their introduction. A GAN comprises two network modules: a generator which maps a sample from a random uniform distribution into an image space, and a discriminator which predicts an image to be real (from the database) or fake (from the generator). Both modules are trained with conflicting objectives based upon the principles of game theory. Compared with the previous approaches which were mainly based upon minimizing pixel-wise reconstruction errors, discriminators in GANs learn a rich similarity metric to discriminate images from non-images. GANs can therefore generate promising and sharper images [25]. GANs are however unstable to train and the generated images are prone to being noisy and incomprehensible. Since their release, many efforts have been made to improve GANs. Radford *et al*. [25] were the first to incorporate a convolutional architecture in GANs, which resulted in improved quality of the generated images. Incorporating side information (class labels) into the discriminator module of GANs has also been shown to improve the quality of generated images [23].

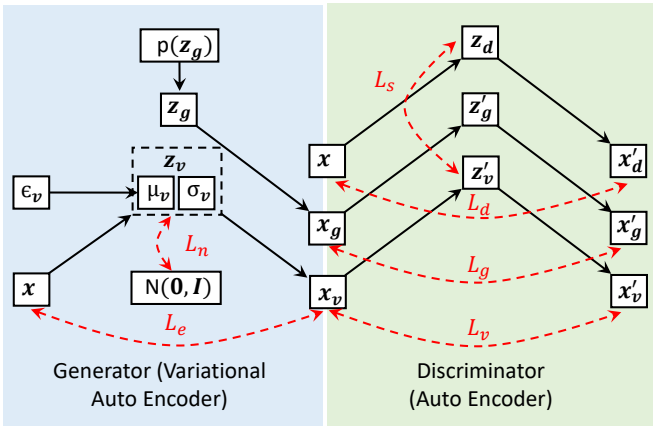
GANs still face many challenges: they are difficult to train and require careful hyper-parameter selection. While training, they can easily suffer from modal collapse [5], a failure mode in which they generate only a single image. Further, it is quite challenging to best balance the convergence of the discriminator and the generator, since the discriminator often easily wins at the beginning of training. Salimans *et al*. [26] proposed architectural improvements that stabilize the training of GANs. In order to strike a balance between the generator and discriminator, Boundary Equilibrium GANs (BEGAN) [4] recently introduced an equilibrium mechanism. BEGANs gradually change the emphasis of generator and discriminator loss terms in gradient descent as the training progresses. Zhao *et al*. proposed Energy based GANs (EBGANs) [33] which model the discriminator as an energy function, and implement it as an auto-encoder. The discriminator energy function can be viewed as a trainable loss function for the generator. EBGANs are more stable and less prone to hyper-parameter selection. EBGANs and earlier GAN versions however lack a mechanism to estimate convergence of the trained model. More recently, Wasserstein GANs (WGANs) [2] introduced a loss that shows a correlation between discriminator loss and perceptual quality of the generated images. Wasserstein loss can therefore act as an indicator for model convergence.

Our work is a continuation of effort to devise stable deep generative models which produce diverse and improved quality images with high resolution. To this end, different from previous works, our proposed model learns to simultaneously match the data distributions, loss distributions and latent distributions of real and fake data during the training stage. We show that such a hierarchical but flexible supervision in the model helps generate images with better quality and high resolution. The closest to our approach is the BEGAN [4] with notable differences including the maximization of a lower bound (Sec. 3.2), perceptual guidance in the latent space (Sec. 3.3), the combination of data and loss distribution matching in the training objective (Sec. 3.4) and incorporation of VAEs in the generator module to learn a latent space where an image of desired style can deterministically be generated using vector arithmetics (Sec. 5.5).

3. Proposed Model

Given a set of data samples $\mathcal{X} = \{\mathbf{x}_i : i \in [1, n]\}$ from an unknown data distribution, we want to learn a generative

Figure 1: **Model Overview:** The red dotted lines represent the loss functions, the downward diagonal arrows (↘) represent the decoding operation and the upwards diagonal arrows (↗) represent the encoding operation. The proposed model first encodes the input image ($\mathbf{x}$) to a parameterized latent representation ($\mathbf{z}_v$) and reconstructs it back using the generator ($\mathbf{x}_g$ and $\mathbf{x}_v$ corresponding to latent representations $\mathbf{z}_g$ and $\mathbf{z}_v$ respectively). The discriminator also consists of an auto-encoder which first encodes the inputs to a latent representation ($\mathbf{z}_d$, $\mathbf{z}'_g$ and $\mathbf{z}'_v$ corresponding to inputs $\mathbf{x}$, $\mathbf{x}_g$ and $\mathbf{x}_v$ respectively), and then reconstructs them back ($\mathbf{x}'_d$, $\mathbf{x}'_g$ and $\mathbf{x}'_v$). The generator and discriminator are trained using the back-propagated error signal from the losses computed to match the actual data, the latent representation and the loss distributions for the real and fake samples.



model with parameters θ that maximizes the likelihood:

$$\theta^* = \max_{\theta \in \Theta} \sum_{i=1}^{n} \log \ell(\mathbf{x}_i; \theta), \tag{1}$$

where $\ell(\cdot)$ denotes the density of each observation $\mathbf{x}_i$. After training, the model can then be used to obtain new samples from the learned distribution.

The proposed model consists of a pair of auto-encoders, as shown in Fig. 1. In contrast to a traditional pixel-wise loss minimization over the data samples, the proposed approach minimizes the loss at the data level, in terms of reconstruction loss and the latent distributions of real and fake images. An input data sample is first encoded and decoded using the CNN blocks which together form a variational auto-encoder whose latent distribution is parameterized as a normal distribution. Afterwards, the reconstructed images are passed through another auto-encoder and the reconstruction error is minimized for both the original and fake images. The overall network is trained using an adversarial loss function which incorporates a learned perceptual similarity metric to obtain high-quality synthetic images. We begin with an overview of the variational auto-encoder, followed by the proposed improvements to enhance the visual quality of generated images.

3.1. Variational Auto-encoder

Auto-encoders prove a very powerful tool to model the relationship between data samples and their latent representations. A variational auto-encoder is similar to the regular auto-encoder in the sense that it first transforms the given data to a low-dimensional latent representation and then projects it back to the original data space. Given an input data sample $\mathbf{x}$, the encoding operation can be represented as: $\mathcal{F}_{vae-e}(\mathbf{x}; \theta_e) = \mathbf{z}_v \sim q_\theta(\mathbf{z}_v|\mathbf{x})$, where $\mathbf{z}_v$ denotes the latent representation. The encoder function $\mathcal{F}_{vae-e}(\mathbf{x}; \theta_e)$ is implemented as a CNN with parameters θ_e. Similarly, the decoding operation is implemented as another CNN and can be represented as: $\mathcal{F}_{vae-d}(\mathbf{z}_v; \theta_d) = \mathbf{x}_v \sim p_\theta(\mathbf{x}|\mathbf{z}_v)$. To be able to reconstruct the original data sample, the following loss function is minimized:

$$\mathcal{L}_e = \frac{1}{|\mathbf{x}|} \parallel \mathbf{x} - \mathcal{F}_{vae-d}(\mathcal{F}_{vae-e}(\mathbf{x}; \theta_e); \theta_d) \parallel_1 \tag{2}$$

Here, $\| \cdot \|_1$ denotes the ℓ_1 norm and $| \cdot |$ denoes the latent space cardinality. The main distinguishing factor of VAE in comparison to a vanilla auto-encoder is the constraint on the encoder to match the low-dimensional latent representation to a prior distribution. The regularization on the encoded latent representation means that the $\mathbf{z}_v$ is constrained to follow a unit Gaussian distribution i.e., $\mathbf{z}_v \sim N(0, I)$. This is achieved by minimizing the Kullback-Leibler (KL) divergence between the two distributions as follows:

$$\mathcal{L}_n = \mathrm{KL}(N(\mu_v, \sigma_v) || N(0, I)) \tag{3}$$

Since an end-to-end training is not possible with the intermediate stochastic step which involves sampling from $N(\mu_v, \sigma_v)$, the re-parametrization trick proposed by [18] is used to enable the error back-propagation by treating the stochastic sampling as an input to the network. As a result, the latent variable is defined as:

$$\mathbf{z}_v = \mu_v + \epsilon_v * \sigma_v, \quad \epsilon_v \sim N(0, I), \tag{4}$$

where μ_v and σ_v denote the mean and variance while ϵ_v is randomly sampled from a unit Gaussian distribution. After the training process, the decoder can be used independently to generate new data samples $\mathbf{x}_g$ by feeding randomly generated samples from the distribution: $\mathbf{z}_g \sim p(\mathbf{z}_g) = N(0, I)$. The main problem with the VAE based models is that the per-pixel loss function $\mathcal{L}_e$ (Eq. 2) defined in terms of mean error results in over-smoothed and blurry images.

To overcome this problem, we propose to match the loss and latent distributions of real and fake images in addition to the minimization of commonly used mean per-pixel error measures. In the following, we first describe the loss distribution matching process and then outline the matching of a perceptual similarity metric using the latent distributions.

3.2. Minimizing Loss Distributions

The VAE set-up described above outputs a reconstructed data sample. The decoding stage in the VAE is similar to the generator function in a regular Generative Adversarial Network (GAN). The generator in a vanilla GAN is followed by a discriminator which distinguishes between the real and fake (generated) data samples. In essence, the distribution of fake and real data samples is matched by playing a game between the generator and the discriminator until the Nash equilibrium is reached. Inspired by Berthelot *et al.* [4], in addition to directly matching the data distributions, we also minimize the approximate Wasserstein distance between the reconstruction loss distributions of real and fake data samples. For sufficiently large number of pixels, the distributions will be approximately normal. The generator will be trained to minimize the reconstruction loss of the fake samples, thus trying to produce real looking samples so that the discriminator assigns them a lower energy. On the other hand, the discriminator will be trained to minimize the reconstruction error for real samples, but maximize the error for the fake samples coming from the generator. We can represent the generator and discriminator loss functions as:

$$\mathcal{L}_{dis} = \mathcal{L}_d - (\mathcal{L}_g + \alpha \mathcal{L}_v) \tag{5}$$

$$\mathcal{L}_{gen} = \mathcal{L}_g + \alpha \mathcal{L}_v \tag{6}$$

where the individual loss terms are defined as:

$$\mathcal{L}_d = \frac{1}{|\mathbf{x}|} \parallel \mathbf{x} - \mathbf{x}'_d \parallel_1, \quad \mathcal{L}_g = \frac{1}{|\mathbf{x}_g|} \parallel \mathbf{x} - \mathbf{x}'_g \parallel_1,$$
$$\mathcal{L}_v = \frac{1}{|\mathbf{x}_v|} \parallel \mathbf{x}_v - \mathbf{x}'_v \parallel_1 . \tag{7}$$

Here α is a weighting parameter which decides the emphasis on the reconstruction loss of recovered training samples from the VAE and $\mathbf{x}'_d$, $\mathbf{x}'_g$ and $\mathbf{x}'_v$ denote the samples reconstructed by the discriminator auto-encoder corresponding to inputs $\mathbf{x}$, $\mathbf{x}_g$ and $\mathbf{x}_v$ respectively:

$$\mathbf{x}'_{v,g,d} = \mathcal{F}_{ae-d}(\mathcal{F}_{ae-e}(\mathbf{x}_{v,g,-}; \theta_{e'}); \theta_{d'}) \tag{8}$$

such that $\theta_{e'}, \theta_{d'}$ represent the parameters of discriminator encoder and decoder respectively. The above defined model can be understood as an improved version of the energy-based generative network [33], where the reconstruction error represents the energy assigned by the discriminator, with low energies being assigned to the samples close to the real data manifold.

We use a simple auto-encoder to estimate the loss distributions of real and fake data. This greatly stabilizes the generative training and avoids high sensitivity to hyper-parameters and undesirable learning modes such as the model collapse. Note that until now, we are matching the reconstructed output from the generator and the loss distributions of real and fake images. This can lead to the generation of simple images which are easy to reconstruct by the discriminator. It is also important to encourage the generator to produce more complex, diverse and realistic samples. For this purpose, we propose to learn a perceptual metric which forces the generator to create more realistic and diverse images.

3.3. Perceptual Guidance

In order to enhance the visual quality of generated images, we propose to add perceptual guidance in the latent space of the discriminator while training the output energy function. This acts as a regularizer during model training and enforces similarity between the real and generated samples using a learned metric. Assuming that the learned latent representation is a compact and perceptually faithful encoding of the real and generated samples, we enforce a loss term in the encoder and generator modules which measures the similarity as an ℓ_1 norm of the difference of the latent representations in the discriminator corresponding to the real and fake images, denoted by $\mathbf{z}_d$ and $\mathbf{z}'_v$ respectively:

$$\mathcal{L}_s = \frac{1}{|\mathbf{z}_d|} \parallel \mathbf{z}_d - \mathbf{z}'_v \parallel_1 . \tag{9}$$

This loss essentially encourages the generator to output images which are close to the data manifold of real images. This is achieved by measuring the similarity between fake and real images in a more abstract sense which aims to roughly match the style of the two image types. Note that we also include a content loss in the the generator training objective to directly minimize the reconstruction error in the VAE model (Eq. 2). The combination of the loss computed in the latent space of the discriminator and the content loss in the VAE model complement each other and result in an optimal training of the model. We describe our adversarial training approach to train the overall model in the next section.

3.4. Adversarial Training

The discriminator in Eq. 5 contains two objectives i.e. to accurately reconstruct real images (formulated as the loss $\mathcal{L}_d$) and to distinguish between the real and fake samples (by maximizing the distance between $\mathcal{L}_d$ and $\mathcal{L}_g$). It is necessary to keep the right balance between these two objectives for an optimal model training. The boundary equilib-

rium technique in [4] used a proportional controller to maintain this balance using a feedback signal. This feedback signal is defined in terms of the error between the weighted reconstruction loss of real and fake data samples. It maintains the equilibrium such that the expected values of both losses are balanced by a constant factor, termed as the diversity factor:

$$\eta = \frac{\mathbb{E}[\mathcal{L}_g] + \alpha\mathbb{E}[\mathcal{L}_v]}{\mathbb{E}[\mathcal{L}_d]} \tag{10}$$

In practice, the proportional control mechanism slowly attains equilibrium when the proportional gain factor is too low. Setting the gain to a higher value leads to unstable learning procedure. Furthermore, since the update is driven by the error signal calculated using the feedback, an exact equilibrium is never attained during the training rather a steady state error is maintained. To avoid these problems, we introduce the accumulated error signal to accelerate the attainment of equilibrium and to avoid the steady state error. We noticed that the accumulated error from the feedback signal can easily cause oscillations around the reference signal. This causes instability during the training and results in suboptimal generated samples. To overcome this problem, we also introduce a differential term while updating the equilibrium parameter 'k_t'. It dampens the oscillations around the reference signal without affecting the signal matching time. In order to stabilize the system, we obtain the derivative term using multiple error samples and use a small differential gain in the update equation as follows:

$$e_t = \eta\mathcal{L}_d - \mathcal{L}_g + \alpha\mathcal{L}_v \qquad \text{at iteration } t \tag{11}$$

$$k_t = k_{t-1} + \lambda_1 e_t + \lambda_2(e_t - e_{t-1}) + \lambda_3(e_t + e_{t-2} - 2e_{t-1}) \tag{12}$$

where, e_t denote the error term, λ_{1-3} denote the gain parameters for the integral, proportional and differential components. The overall loss function can therefore be expressed as:

$$\mathcal{L}_{dis} = \mathcal{L}_d - k_{t-1}(\mathcal{L}_g + \alpha\mathcal{L}_v) \tag{13}$$

$$\mathcal{L}_{gen} = \mathcal{L}_g + \alpha\mathcal{L}_v + \beta\mathcal{L}_s + \gamma\mathcal{L}_e \tag{14}$$

$$\mathcal{L}_{enc} = \mathcal{L}_n + \beta\mathcal{L}_s + \gamma\mathcal{L}_e \tag{15}$$

where, α, β, γ denote the weights which put different level of emphasis on the respective loss functions. Note that our generator and discriminator loss functions incorporate the reconstruction loss $\mathcal{L}_v$ computed on the real data samples reconstructed by the VAE. This forces the generator to produce high quality samples lying close to the real data manifold. Similarly, the generator model is trained on the data loss ($\mathcal{L}_e$), the reconstruction loss and the latent space similarity loss ($\mathcal{L}_s$). The convergence rate of the model can be measured by analyzing the error measure given by $\mathcal{L}_d + |\eta\mathcal{L}_d - \mathcal{L}_g - \alpha\mathcal{L}_v|$. The overall training is illustrated in Algorithm 1.

Algorithm 1: Learning procedure for proposed model

Initialization: $\theta_e, \theta_d, \theta_{e'}, \theta_{d'} \leftarrow$ randomly initialize VAE and AE encoder, decoder respectively.
// Perform a total of T training iterations
for $t = 1 : T$ **do**
  VAE Training
  $\mathbf{X} \leftarrow$ sample a random batch from $\mathcal{X}$
  $\mathbf{Z}_v \leftarrow \mathcal{F}_{vae-e}(\mathbf{X}; \theta_e)$
  $\mathbf{Z}_g \sim N(\mathbf{0}, \mathbf{I})$
  $\mathbf{X}_{v,g} \leftarrow \mathcal{F}_{vae-d}(\mathbf{Z}_{v,g}; \theta_d)$
  $\mathbf{Z}'_{v,g,d} \leftarrow \mathcal{F}_{ae-e}(\mathbf{X}_{v,g,-}; \theta_{e'})$
  Calculate $\mathcal{L}_e$, $\mathcal{L}_n$ and $\mathcal{L}_s$ using Eqs. 2, 3 and 9
  $\theta_e \leftarrow \nabla_{\theta_e}\mathcal{L}_{enc}$ (Eq. 15)
  Adversarial Training
  $\mathbf{X}'_{v,g,d} \leftarrow \mathcal{F}_{ae-d}(\mathbf{Z}'_{v,g,d}; \theta_{d'})$
  Calculate $\mathcal{L}_g, \mathcal{L}_d$ and $\mathcal{L}_v$ using Eq. 7
  $\theta_d \leftarrow \nabla_{\theta_d}\mathcal{L}_{gen}$ (Eq. 14)
  Calculate η, e_t and k_t using Eqs. 10, 11 and 12
  $\theta_{e'}, \theta_{d'} \leftarrow \nabla_{\theta_{e'}, \theta_{d'}}\mathcal{L}_{dis}$ (Eq. 13)
Return: Updated parameters $\theta_e, \theta_d, \theta_{e'}, \theta_{d'}$

4. Implementation Details

Both the generator and discriminator in our proposed model consist of an encoder and a decoder. For simplicity, we keep the backbone architecture of both the generator and the discriminator identical, except that the generator models a VAE. With in the generator and the discriminator, the architecture of encoder and decoder are also equivalent to each other in terms of the number of layers and therefore the parameters. We keep the design of encoder and decoder consistent with the discriminator module of [4] This design choice was made due to two reasons: **(1)** The encoder and decoder architectures are fairly simple compared to their counterpart models used in GANs. These consist of three pairs of convolution layers, each with an exponential linear unit (ELU) non-linearity and interleaved with sub-sampling and up-sampling layers for the case of encoder and decoder respectively. Furthermore, the architecture does not use dropout, batch-normalization and convolution transpose layers as for the case of other competing model architectures. **(2)** It makes it easy to compare our model with the latest state of the art BEGAN model [4], which achieves good results in terms of visual quality of the generated images. Note that different from [4], our generator module is based upon a VAE and both the generator and the discriminator takes into account data, loss and latent distributions of real and fake data.

The latent vector $\mathbf{z} \in \mathbb{R}^N$ is sampled from $[0, 1]$ following a normal distribution. During training, the complete model shown in Fig. 1 is used. At test time, the encoder and

Figure 2: Generated Images (64×64)

Figure 3: Generated Images (128×128)

the discriminator are discarded and random samples from $p(\mathbf{z}_g)$ are feed forwarded through generator to obtain new data samples. We use the gain parameters settings to be $10^{-3}, 10^{-5}$ and 10^{-5} respectively in our experiments. The parameters α, β, γ and η are set to $0.3, 0.1, 0.1$ and 0.5 respectively. The model was trained for 300k iterations using the Adam optimizer initialized with a small learning rate of 5×10^{-5}.

5. Experiments

5.1. Dataset

We use the CelebFaces Attribute Dataset (CelebA) [21] in our experiments. CelebA is a large scale dataset with more than 202k face images of over 10k celebrities. The dataset contains a wide range of appearances, head poses and backgrounds. The dataset has been annotated with 40 attributes, however, these are not used in our unsupervised training of the proposed model. We use the aligned and cropped version of images in the dataset where the face appears roughly at the center of an image.

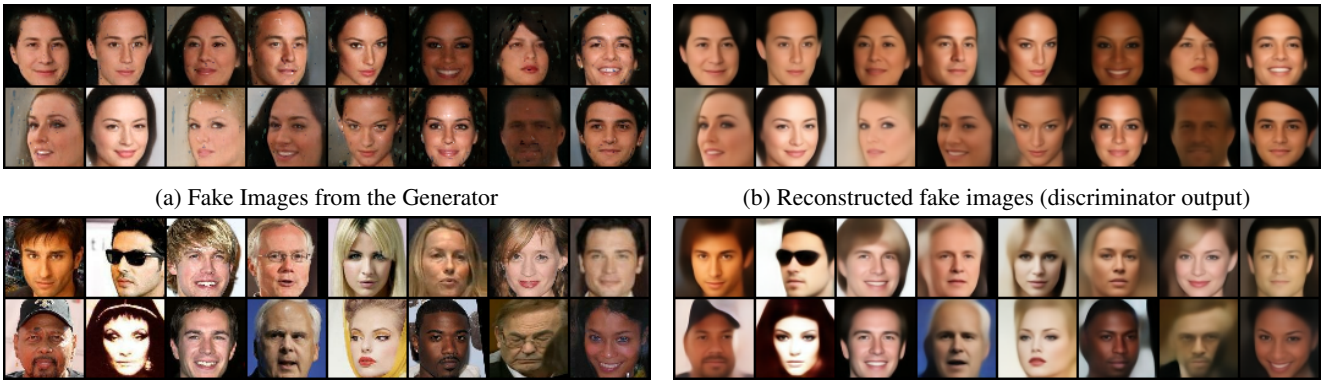
5.2. Results

Our sample generated face images are shown in Figures 2, 3 and 4. Our approach was able to generate a diverse set of face images, belonging to different age groups and containing a variety of attributes such as the blonde hair, smiling face and heavy make-up. The generated images contain both male and female examples, although we noticed a bias towards the female category due to a heavy representation of female celebrities in the celebA dataset. The generated samples also contain pose variations and a diverse set of facial expressions.

Approach	Inception Score
DCGAN (ICLR'16) [25]	4.89
Improved GAN (NIPS'16) [26]	4.36
ALI (ICLR'17) [5]	5.34
MIX + WGAN (ICML'17) [3]	4.04
PixelCN++ (ICLR'17) [27]	5.51
AS-VAE-g (NIPS'17) [24]	**6.89**
BEGAN (Arxiv'17) [4]	5.62
Ours (BEGAN* + LSM)	6.12
Ours (VAE + BEGAN* + LSM)	**6.80**
Improved GAN (semi-supervised) [26]	8.09
Real Images	11.24

Table 1: Quantitative comparison on the CIFAR-10 dataset in terms of Inception score. The best and the second best performances are shown in red and blue respectively. BEGAN* denotes [4] with the modified equilibrium approach and LSM stands for the Learned Similarity Metric. All the reported performances are for unsupervised cases except the bottom two, which use label information or real images, respectively.

The proposed method allowed us to easily scale up the resolution of generated samples while maintaining their visual quality (see Figure 3). However, we noticed some smoothing effects when the model was trained to generate 128×128 images compared to 64×64 images. Since the discriminator in our model is an auto-encoder, it is also of interest to visualize the reconstruction performance for both the real and fake images (Figure 4). Such a reconstruction is not possible with the original GAN based model. We noticed that the generated samples are less complex, and are therefore reconstructed very well by the discriminator. In



(a) Fake Images from the Generator

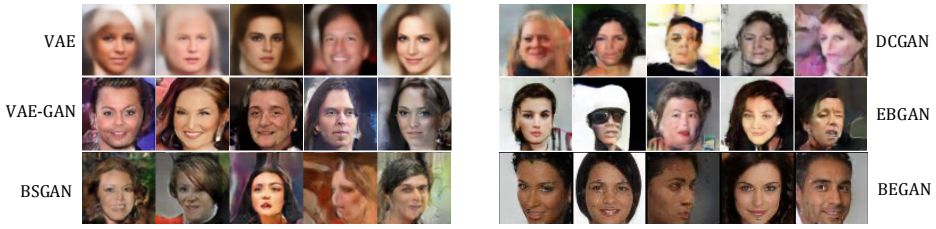
(b) Reconstructed fake images (discriminator output)

(c) Real Images

(d) Reconstructed real images (discriminator output)

Figure 4: Real and fake images and their corresponding reconstructions by the discriminator.

Figure 5: Qualitative comparison of face generation results with other recent image generation approaches.



contrast, the real images are more challenging to reproduce and therefore suffer from over-smoothing by the generator. It also shows that the equilibrium between generator and discriminator is maintained till the end of the training process.

5.3. Quantitative Comparisons

For quantitative evaluation of our approach, we compute the Inception score proposed in [26] for the CIFAR-10 dataset. The training procedure is carried out in an unsupervised manner for a fair comparison with the previous techniques listed in Table 1. Precisely, for every generated image, the Inception model [28] is used to calculate $p(y|\mathbf{x})$, where y denotes the class label. Good quality samples are expected to have $p(y|\mathbf{x})$ with low entropy (i.e., consistent predictions) and the marginalized distribution over all samples $p(y)$ with high entropy (i.e. diverse predictions). The inception score is defined as the combination of these two criterion: $\exp(\mathbb{E}[\mathrm{KL}(p(y|\mathbf{x}) \parallel p(y))])$. This measure was found to match well with the human evaluations of the quality of generated samples. Our quantitative results show the significance of using a learned similarity metric which provides significant improvement over the BEGAN [4] approach. The proposed model with the inclusion of VAE with additional constraints on the latent representations lead to the significantly better performance compared to the closely related BEGAN. For comparisons, we also report the oracle case with real images, and the case where semi-supervised learning is performed as proposed in [26]. Our unsupervised model scores fairly close to the semi-supervised version of [26].

5.4. Qualitative Comparisons

We qualitatively compare our results with: **(i)** a VAE model comprising of an encoder and a decoder [18], **(ii)** a deep convolutional GAN [25], **(iii)** combined VAE+GAN model [19], **(iv)** energy based GAN [33], **(v)** boundary equilibrium GAN [4] and **(vi)** the boundary seeking GAN [14]. Note that the results reported in [4] are obtained by training on a much larger (but publicly unavailable) dataset comprising of 360k celebrity face images. For a fair comparison, we train their model on the CelebA dataset with the same parameter settings as reported by the authors. All other approaches were already trained on the CelebA dataset, therefore we take their reported results for comparison. Since the results for larger image sizes are not available for most of these approaches, we compare for the case of 64×64 images (see Figure 5).

The VAE model implemented using the convolutional encoder and decoder is trained using the pixel wise loss and therefore generates blurry images with high bias towards the frontalized head poses. The deep convolutional GAN and energy based GAN generate much sharp images but contain both local and global visual artifacts. The combination of VAE and GAN gives better results than the original VAE based model, however it still does not solve the blur and

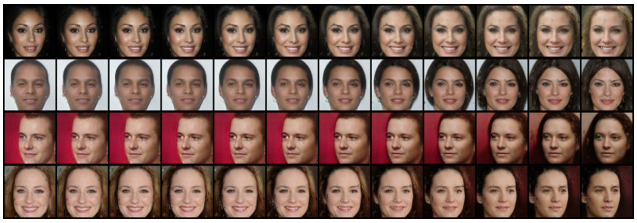


Figure 6: Interpolation between the two generated images (64×64) by moving in the latent space.

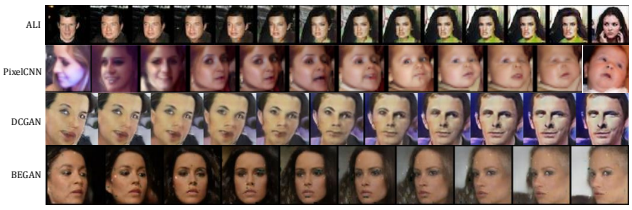


Figure 7: Comparison for continuity in latent space.

noise in the output images. The boundary equilibrium GAN generates fairly good quality images with a diverse set of facial attributes, however their model have problems dealing with side poses. Lastly, the boundary seeking GANs have high contrast close to face boundaries but the overall faces look unrealistic.

5.5. Exploring the Latent Space

In order to validate that the learned generator has not merely memorized the training examples, we experiment the continuity in the latent space to check if the intermediate latent representations also correspond to realistic images. To this end, we find the latent vectors z corresponding to two real images. We then interpolate between the two z embeddings and show the corresponding images generated by the model. The results are reported in Figure 6. We note that there exists smooth transition between the real faces, even in cases where the two images are remarkably different. This proves that the model has good generalization capability and has not memorized the image contents. Here, we also qualitatively compare with other image generation methods including ALI [5], PixelCNN [29], DCGAN [25] and BEGAN [4] in Figure 7.

A distinguishing feature of our model is the flexibility to control the generated images by appropriately modifying inputs in the latent space. This is in contrast to regular GAN based models, which do not allow such a control over the latent space since the samples are randomly drawn from a uniform distribution. In this experiment, we aim to study the relationships between the latent vectors and the face attributes. For this purpose, we encode each image using the VAE and average the latent vectors to obtain a representation for each of the 40 semantically meaningful attributes

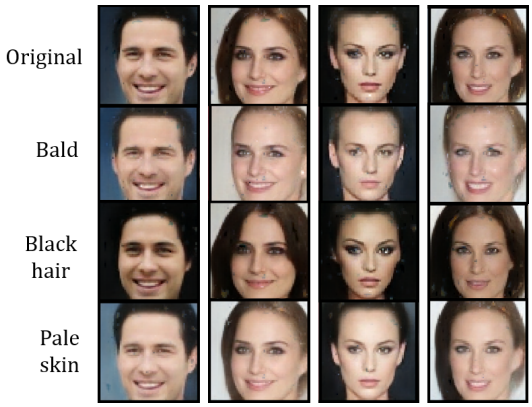


Figure 8: Latent space arithmetic results for example attributes.

in the celebA dataset. Similar to [22], we show that applying simple arithmetic operations in the latent space using the average response corresponding to each attribute results in the modified images with the desired attributes. Specifically, we calculate the average latent encoded representations of all images with and without a specific attribute respectively. Then a weighted version of the difference vector can be added to the newly generated image latent representation to add or remove a specific attribute. We show some examples of latent space arithmetic in Figure 8 where attributes such as bald, black hair and pale skin are clearly reflected on to the generated images.

6. Conclusion

This paper proposed a new approach to train VAEs by matching the data as well as the loss distributions of the real and fake images. This was achieved by a pair of auto-encoders which served as the generator and the discriminator in the adversarial training. Our model automatically learned a similarity metric defined in terms of latent representation obtained using the discriminator to enable generation of high-quality image outputs. This helped in overcoming the artifacts caused when only the data or loss distribution was matched between samples. Our method utilized a simple model architecture, was stable during training and easily scaled up to generate high dimensional perceptually realistic outputs. In future, we will explore the conditional image generation models to obtain even higher image resolutions while maintaining the photo-realistic quality.

Acknowledgments

Thanks to NVIDIA Corporation for donation of the Titan X Pascal GPU used for this research.

References

[1] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *NIPS 2016 Workshop on Adversarial Training. In review for ICLR*, volume 2016, 2017. 1

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 1, 2

[3] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573*, 2017. 6

[4] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 1, 2, 4, 5, 6, 7, 8

[5] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016. 2, 6, 8

[6] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1033–1038. IEEE, 1999. 2

[7] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002. 2

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2

[9] M. Hayat, S. H. Khan, and M. Bennamoun. Empowering simple binary classifiers for image set based face recognition. *International Journal of Computer Vision*, 2017. 1

[10] M. Hayat, S. H. Khan, M. Bennamoun, and S. An. A spatial layout and scale invariant feature representation for indoor scene classification. *IEEE Transactions on Image Processing*, 25(10):4829–4841, 2016. 1

[11] J. Hays and A. A. Efros. Scene completion using millions of photographs. In *ACM Transactions on Graphics (TOG)*, volume 26, page 4. ACM, 2007. 2

[12] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 2

[13] G. E. Hinton and T. J. Sejnowski. Learning and releaming in boltzmann machines. *Parallel Distrilmted Processing*, 1, 1986. 2

[14] R. D. Hjelm, A. P. Jacob, T. Che, K. Cho, and Y. Bengio. Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*, 2017. 7

[15] S. H. Khan, M. Hayat, M. Bennamoun, F. Sohel, and R. Togneri. Cost sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 2017. 1

[16] S. H. Khan, M. Hayat, M. Bennamoun, R. Togneri, and F. A. Sohel. A discriminative representation of convolutional features for indoor scene recognition. *IEEE Transactions on Image Processing*, 25(7):3372–3383, 2016. 1

[17] S. H. Khan, M. Hayat, and F. Porikli. Scene categorization with spectral features. *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2, 3, 7

[19] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. 1, 7

[20] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016. 1

[21] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 6

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 8

[23] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016. 2

[24] Y. Pu, W. Wang, R. Henao, L. Chen, Z. Gan, C. Li, and L. Carin. Adversarial symmetric variational autoencoder. In *Advances in Neural Information Processing Systems*, pages 4331–4340, 2017. 6

[25] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1, 2, 6, 7, 8

[26] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2226–2234, 2016. 2, 6, 7

[27] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017. 6

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 7

[29] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016. 8

[30] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016. 1

[31] R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2016. 1

[32] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2528–2535. IEEE, 2010. 1

[33] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016. 1, 2, 4, 7